Set characterization-selection towards classification based on interaction index

J. Murillo, S. Guillaume, F. Spetale, E. Tapia, P. Bulacio

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

Set characterization-selection towards classification based on interaction index

J. Murillo[a,*], S. Guillaume[b], F. Spetale[a], E. Tapia[a], P. Bulacio[a]

[a] *CIFASIS-CONICET, Universidad Nacional de Rosario, Argentina.*
[b] *Irstea, UMR ITAP, 34196 Montpellier, France.*

Abstract

In many real world datasets both the individual and coordinated action of features may be relevant for class identification. In this paper, a computational strategy for relevant feature selection based on the characterization of redundant or complementary features is proposed. The characterization is achieved using fuzzy measures and an interaction index computed from fuzzy measure coefficients. Fuzzy measure identification requires raw data to be turned into confidence degrees. This key step is carried out considering the distributions of feature values across all the classes. Fuzzy measure coefficients are then estimated with an improved version of the Heuristic Least Mean Squares algorithm that includes an efficient management of untouched coefficients. Then, a generalization of the Shapley index for an arbitrary number of features is used. Simulations experiments on synthetic datasets are performed to study the behavior of this generalized interaction index. For extreme datasets, containing either redundant or complementary features as well as noise, the index value is defined by mathematical formula. This result is used to motivate feature selection guidelines that take into account feature interactions. Experimental results on benchmark datasets show that the proposal allows for the design of compact, interpretable and competitive classification models.

Keywords: Choquet, Subset characterization, HLMS, Generalized Shapley index.

*Corresponding author
Email address: murillo@cifasis-conicet.gov.ar (J. Murillo)

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

1. Introduction

Feature selection is a critical task in classification problems, especially when dealing with noisy data of complex structure. Common feature selection techniques rely on individual feature evaluations [1] assuming independence between features. However, this condition is rarely met in practice [2, 3]. For example, in micro-array datasets, the expression of certain genes, known to be biologically important for class differentiation, may appear relevant only when evaluated within groups [4]. This suggests that in complex data domains, even if the support of an individual feature may be small, its contribution to the support of feature subsets may be significant.

Diverse techniques have been proposed for evaluating the support of feature subsets in classification problems, including genetic algorithms [5], statistical methods based on mutual information [6] and fuzzy measures [7, 8, 9]. A main drawback of genetic approaches is that they cannot explain the selection criteria; this drawback is partially solved by mutual information approaches able to explain the selection of feature pairs. On the other hand, fuzzy measures naturally characterize the relevance of feature subsets of arbitrary cardinality for general decision making problems. Hence, they are good candidates for tackling the feature selection and classification problem.

In practice, the estimation of fuzzy measures for n features entails the identification of 2^n coefficients. For small values of n, such identification problem may be solved by a domain expert. However, when n gets moderate or large, machine learning procedures are required. Both genetic algorithms [10, 11, 12, 13] and gradient descent approaches [14, 15, 16, 17, 18] have been considered in literature. The Heuristic Least Mean Squares ($HLMS$) algorithm [15, 19, 20], based on the Choquet Integral, is a representative of the gradient methods. They exhibit two important properties, the traceability of fuzzy measure estimation due to the lack of random steps and the efficient use of training data [21]. They also need much less space. While multiple instances of fuzzy measure coefficients are required by genetic algorithms, $HLMS$ uses only one. Furthermore, the $HLMS$ heuristic based on the error function minimization of individual samples, usually leads to less extreme solutions [22]. We note, however, that the original $HLMS$ formulation has some convergence problems and thus, its revised $HLMSr$ version [20] is recommended.

Fuzzy measures weight all the possible combinations subsets. Hence, to absolutely characterize a feature subset we need to consider its support

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

across all subsets. This can be accomplished by computing the generalized interaction index proposed in [23], hereafter called GI. Its computation requires reliable fuzzy measures and thus, special attention must be paid to raw data preprocessing steps and to the fuzzy measures identification process itself.

In this paper, a robust feature selection strategy, based on the GI characterization of feature subsets regarding their relevance to classification problems, is proposed. With this intention, we first analyze GI results when standard raw data conversion procedure is used for the $HLMSr$ estimation of fuzzy measures. As a result of this study, a raw data conversion procedure based on normalized confidence degrees and an improved version of the $HLMSr$ algorithm are proposed. Aiming to understand some unexpected GI results previously reported in literature [14, 23], we study the GI behavior on homogeneous datasets, i.e., datasets where, in each class, all features are either redundant or complementary possibly including noisy features. As a result of this study, some guidelines for the efficient GI characterization of feature subsets in homogeneous datasets are introduced.

The outline of the paper is as follows. In Section 2, fuzzy measures, the discrete Choquet integral and GI are briefly reviewed. Then, fuzzy measure identification process with $HLMSr$ is analyzed, and some modifications to the algorithm are introduced in Section 3. Section 4 deals with the conversion from raw data into confidence degrees. In Section 5, the GI behavior is studied yielding two theorems (proofs are in the appendix). The usefulness of feature subset selection guided by GI characterizations is evaluated in Section 6 with UCI[1] benchmark classification problems. Section 7 summarizes the main conclusions and perspectives.

2. Preliminaries

This section introduces basic concepts related to fuzzy measures, to the discrete Choquet integral and to the GI index. The reader may refer to [8, 9, 24] for further details. Let us consider a finite set $X = \{x_1, \ldots, x_i, \ldots, x_n\}$ and let $\mathcal{P}(X)$ denote its power set.

[1] http://archive.ics.uci.edu/ml/

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

2.1. Fuzzy measures and the discrete Choquet integral

A fuzzy measure (FM) is a set function $\mu\colon \mathcal{P}(X) \to [0,1]$ fulfilling the following two axioms:

1. Normalization: $\mu(\emptyset) = 0$, $\mu(X) = 1$
2. Monotonicity: $A \subseteq B \subseteq X \Rightarrow \mu(A) \leq \mu(B)$

While the former allows for fuzzy measure comparisons, the latter ensures that adding any element to a given subset does not make it less informative.

Fuzzy measures $\mu\colon \mathcal{P}(X) \to [0,1]$ are used in the definition of discrete Choquet integrals. For a given $f\colon X \to \Re^+$, its discrete Choquet integral $\mathcal{C}$ with respect to a fuzzy measure $\mu\colon \mathcal{P}(X) \to [0,1]$ is defined as follows:

$$\mathcal{C}_\mu\big(f(x_1),..,f(x_n)\big) \triangleq \sum_{i=1}^{n} \big(f(x_{(i)}) - f(x_{(i-1)})\big)\mu(A_{(i)}) \tag{1}$$

where $x_{(\cdot)}$ is the rearrangement induced by $f(x_i)$, $i = 1,\ldots,n$, sorted in ascending order, i.e., $f(x_{(1)}) < \cdots < f(x_{(n)})$ and $A_{(i)} = \{x_{(i)}, x_{(i+1)}, \ldots, x_{(n)}\}$; by convention $f(x_{(0)}) = 0$.

2.2. The GI characterization of features subsets

In the field of cooperative game theory, the Shapley index can be used to characterize the importance of individual features [25]:

$$\phi_i = \sum_{K \subseteq X \setminus i} \frac{(n - |K| - 1)!|K|!}{n!}\big(\mu(K \cup \{i\}) - \mu(K)\big) \tag{2}$$

where $|\cdot|$ indicates the cardinality, and 0!=1 as usual. The Shapley value of μ is the vector $\phi = [\phi_1 \cdots \phi_n]$ which has the property to be linear with respect to μ, and to satisfy:

$$\sum_{i=1}^{n} \phi_i = \mu(X) = 1 \tag{3}$$

This index has been generalized, first to characterize the importance of feature pairs [26] and finally subsets of arbitrary cardinality [23]:

$$GI(A) = \sum_{K \subseteq Y \setminus A} \xi(K) \sum_{B \subseteq A} (-1)^{|A|-|B|}\mu(K \cup B) \tag{4}$$

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

where

$$\xi(K) = \frac{(n - |K| - |A|)!|K|!}{(n - |A| + 1)!}$$

Note that $GI(A)$ reduces to the Interaction index when only two elements belong to A and it further reduces to Shapley index when A is a singleton. While the Shapley index is known to range in $[0, 1]$ and the Interaction index in $[-1, 1]$, the problem of characterizing the GI range for feature subsets of arbitrary cardinality remains open.

3. Improving *HLMSr* for reliable fuzzy measure identification

To accomplish feature characterization and selection based on GI, a reliable fuzzy measure estimation is needed. A promising alternative is the improved version of $HLMS$ [15] proposed by the authors in [20] called $HLMSr$. Briefly, $HLMSr$ is an iterative supervised gradient-based algorithm for the identification of fuzzy measures. The algorithm starts from a set of m training samples involving n features and a target function T. At the beginning, fuzzy measures coefficients $\mathbf{u}$ are initialized to the so-called equilibrium state [15]. This initialization strategy reduces the Choquet integral to a simple arithmetic mean. At each iteration step $HLMSr$ updates the FM coefficient values, for each sample x^j with $j = 1, ..., m$, according to the difference between the target T^j and the current Choquet integral (Line 12). A given sample always uses the same coefficients to compute the integral, one for each subset size between 1 and $n - 1$. The coefficient associated to a subset of size l is called u_l.

The revised version used in this paper counts with some improvements with respect to the initial version [15]. Firstly, the update formula (Line 12) has been modified to get a true gradient and allows for coefficients to converge to the expected value with synthetic data sets. This modification is discussed in detail in [20]. Secondly, the management of untouched coefficients is changed. There might be coefficients which do not participate in any integral computation. As their initial value may limit the evolution of neighboring coefficients, they are identified (Line 6), and not taken into consideration for monotonicity checking (Line 13). The value of an untouched coefficient u_l is set at the end of the algorithm (Line 18). The values fulfilling the monotonicity condition verify Eq.(5):

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

$$max\{u_{(l-1)}\} \leq u_l \leq min\{u_{(l+1)}\} \tag{5}$$

These coefficients are set to the minimal authorized value in order not to influence GI values. Finally, $HLMSr$ proved sensitive to the default lexicographic order of equal valued features used in the computation of the core Choquet integral. In this case, only the coefficient of the first feature is updated. Since there is no reason to use a lexicographic order of features values, equal feature values are randomly ordered. As a result, features providing the same information are now considered equally relevant.

Algorithm 1 Revised HLMS

Input: Training dataset D with samples $(\mathbf{x}^j, T^j)$, $j = 1, ..., m$
Output: Fuzzy measure coefficients $\mathbf{u} = [u_p], p \in \mathcal{P}(X)$
for $p \in \mathcal{P}(X)$ **do** {Initialization}
  $u_p = \frac{|p|}{|X|}$
end for
Identify_Untouched(D)
repeat
  examples $\leftarrow$ random$(1 : m)$ {Sensitivity to data presentation order}
  for $j \in$ examples **do**
    $e^j = \mathcal{C}_u(\mathbf{x}^j) - T^j$ {Individual error calculation}
    for $l \in (1 : n-1)$ **do**
      $u_l = u_l - \alpha \times \frac{e^j}{e_{max}} \times (x^j_{(n-l+1)} - x^j_{(n-l)})$ {Coefficient update}
      Neighbors_Monotonicity (u_l) {Monotonicity check}
    end for
  end for
  $E \leftarrow \sqrt{\frac{1}{m}\sum_{j=1}^{m}(\mathcal{C}_u(\mathbf{x}^j) - T^j)^2}$ {Error calculation}
until Stop_Criterion(E)
Untouched_Monotonicity(D, $\mathbf{u}$) {Final monotonicity correction}

The inputs of $HLMSr$ are commensurable confidence degrees. The first step of the feature characterization process has to transform raw feature values into confidence degrees.

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

4. Raw data to confidence degree conversion

Raw data cannot be used for fuzzy measure identification using $HLMSr$. Indeed, what is needed is commensurable information, i.e., the values must be in the same scale and have a common meaning, for instance the higher the more likely to belong to the class (a detailed discussion can be found in [27]). Many methods for raw data conversion have been proposed in literature, including Parzen Windows, possibilistic histograms and Gaussian densities [21]. Among these proposals, Gaussian densities appear as a good option due to their conceptual simplicity and low computational complexity. They have been used, e.g. for image features aggregation [28].

Using Gaussian densities, the relevance of an individual feature value for the identification of a given class is taken from the probability density distribution of feature values within the class. The transformed feature value is then interpreted as a partial evidence that the corresponding sample belongs to the given class: the higher the transformed feature value, the higher the evidence provided by the feature that the sample belongs to the given class. Overall sample support to a given class can be then obtained by the aggregation of partial evidences, e.g., by means of the discrete Choquet integral. Though useful and intuitive, the Gaussian transformation exhibits some problems when dealing with feature values falling in distribution tails. For example, let us consider the synthetic classification problem shown in Figure 1. This problem involves two relevant features (V1 and V2) and one noisy feature (V3). For class 1, features V1 and V2 are complementary, since both are needed to classify the samples. On the other hand, they are redundant for class 2. Although these observations are acceptable reflected by GI results shown in Table 4, $HLMSr$ classification results are poor: a detailed inspection of misclassified samples shows that the problem is at the small relevance assigned to feature values falling in the Gaussian tails.

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

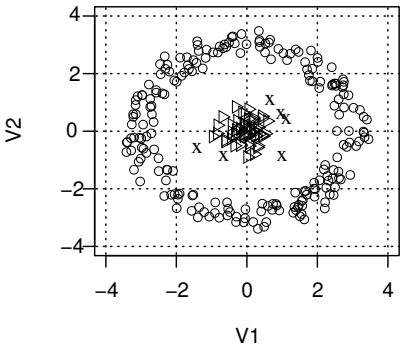

Figure 1: The synthetic *Atom* dataset in the V1-V2 projection. Class 1 is plotted with triangles (nucleon) and class 2 with circles (ring). Misclassified samples are marked with crosses.

Coalition	Class 1	Class 2
V1	0.52	0.22
V2	0.33	0.69
V3	0.14	0.07
V1-V2	0.61	-0.41
V1-V3	0.15	0.06
V2-V3	-0.24	-0.11

Table 1: *GI* results for the *Atom* dataset. For class 1, the *GI* indicates that both V1 and V2 are relevant and quite complementary. For class 2, the *GI* indicates that feature V2 is more relevant than V1 and that they are quite redundant.

Actually, what is happening is that some transformed feature values are being underestimated, i.e., although the absolute relevance of a feature value for a given class may be low, it should be consider relevant if being absolutely irrelevant for remaining classes. This observation suggests that raw data conversion processes must also take into account the distribution of features across all classes. Hence, the following process to convert raw data into confidence degrees is proposed:

$$\mathring{x}_{ik}^{j} = \frac{\tilde{x}_{ik}^{j}}{\{\sum_{s=1}^{p} \tilde{x}_{is}^{j}\}} \tag{6}$$

where $\tilde{x}_{is}^{j}$ is the frequency of the x^j raw value in the Gaussian distribution of feature i for class s. $\mathring{x}_{ik}^{j}$ represents the normalized confidence degree over all the classes. This normalization remarkably improved classification results (100% of accuracy for the *Atom* dataset). However, anomalous *GI* results were still observed for completely redundant features, i.e., with features having exactly the same values. Although, these type of features rarely occur in practice, they are frequent after raw data conversion into confidence degrees.

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

Let us consider the data shown in Figure 2. The problem involves two completely relevant redundant features ($V1$ and $V2$) as well as a noisy one, $V3$, not shown). For both classes, any of the $V1$ or $V2$ features is able to classify the samples. This leads to binary confidence degrees, either $\{1, 1\}$ or $\{0, 0\}$ for the pair $V1$-$V2$. In this case only the pair, and none of the singletons, would be updated, yielding unexpected GI values.

This problem is solved by introducing a small amount of random noise, $\varepsilon = \mathcal{N}(0, 0.01)$:

$$\mathring{x}_{ik}^{j} = \frac{\tilde{x}_{ik}^{j}}{\{\sum_{s=1}^{p} \tilde{x}_{is}^{j}\} + \varepsilon} \tag{7}$$

Thanks to the noise introduction, the GI results, shown in Table 4, are the expected ones and the classification is correctly achieved.

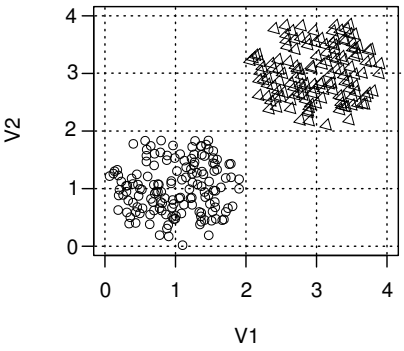

Figure 2: Synthetic *Balls* dataset with the V1-V2 projection. Class 1 is plotted with triangles (top) and class 2 with circles (bottom).

Coalition	Class 1	Class 2
V1	0.5	0.5
V2	0.5	0.5
V3	0	0
V1-V2	-1	-1
V1-V3	0	0
V2-V3	0	0

Table 2: GI results for the *Balls* dataset. For class 1, the GI indicates that both features V1 and V2 relevant and that they are completely redundant.

This study highlights the importance of this preprocessing step. Now that fuzzy measures seem to be correctly identified, the next step is to analyze the GI behavior.

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

5. Characterization of feature subsets by fuzzy measures and the *GI* index

To gain insight into the GI behavior, simulations were performed using a synthetic *Master* dataset with seven normalized and commensurable features, i.e., normalized confidence degrees, $A1$ to $A7$, and one binary target T (see Table 3). It can be observed that features $A1$ and $A2$ are redundant for class 1 (T=1) since either $A1$ or $A2$ can be used for class identification. Features $A3$ and $A4$ are complementary for class 1. Finally, features $A5$ to $A7$ can be considered noise for both classes. From the *Master* data, three datasets, Inf, *Red* and $Comp$. They correspond to typical situations like noise, redundancy or complementariness. The Inf dataset, $A1$, $A5$ to $A7$ and T, contains only one informative feature, $A1$. The *Red* one, $A1$, $A2$, $A5$, $A6$ and T, includes two redundant features, $A1$ and $A2$. In the $Comp$ dataset, $A3$, $A4$, $A5$, $A6$ and T, $A3$ and $A4$ are complementary for class 1. These three datasets are used to analyze the effectiveness of $HLMSr$ for fuzzy measure estimation and the posterior computation of GI values. In these experiments, $HLMSr$ was set to work with a maximum of 1000 iterations and a learning rate $\alpha = 0.01$.

Sample	$A1$	$A2$	$A3$	$A4$	$A5$	$A6$	$A7$	T
1	0.60	0.80	0.80	0.90	0.55	0.33	0.53	1
2	0.60	0.55	0.80	0.80	0.97	0.01	0.29	1
3	0.70	0.90	0.75	0.80	0.72	0.27	0.17	1
4	0.60	0.80	0.70	0.60	0.63	0.30	0.56	1
5	0.75	0.90	0.80	0.80	0.04	0.68	0.05	1
6	0.65	0.60	0.90	0.90	0.00	0.47	0.03	1
7	0.80	0.75	0.70	0.80	0.96	0.65	0.07	1
8	0.70	0.90	0.80	0.70	0.96	0.81	0.16	1
9	0.00	0.10	0.10	0.50	0.41	0.17	0.28	0
10	0.10	0.20	0.40	0.10	0.78	0.35	0.46	0
11	0.20	0.30	0.50	0.15	0.17	0.64	0.57	0
12	0.10	0.20	0.10	0.60	0.45	0.73	0.18	0
13	0.20	0.10	0.50	0.20	0.48	0.68	0.47	0
14	0.30	0.35	0.60	0.10	0.08	0.49	0.60	0
15	0.10	0.20	0.70	0.20	0.65	0.92	0.04	0
16	0.20	0.25	0.10	0.80	0.61	0.25	0.50	0

Table 3: The synthetic *Master* dataset after relative raw data conversion across 16 samples. Columns $A1$ to $A7$ correspond to features values and column T to the target value.

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

Results on the Inf dataset are shown in Table 4. Fuzzy measure coefficients turn to be binary valued, being one for the informative feature $A1$, and for all subsets containing it, in agreement with the monotonicity constraint. As expected, they are zero for all subsets including only noisy features. Although being informative, fuzzy measure coefficient may be difficult to analyze due to the monotonicity constraint. The GI values make the analysis easy: in this case, the only non null value is the one of the informative feature.

FM set	A1	A5	A6	A7	A1-A5	A1-A6	A1-A7	A5-A6	A5-A7	A6-A7	A1-A5-A6	A1-A5-A7	A1-A6-A7	A5-A6-A7
FM value	1	0	0	0	1	1	1	0	0	0	1	1	1	0
GI	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4: Inf dataset: Fuzzy measure (FM) coefficients and GI values

Results on the *Red* dataset are shown in Table 5. As expected, Fuzzy measure coefficients are zero for all combinations of noisy features. For subsets containing features $A1$ or $A2$ they seem to indicate the relevance of these features but do not directly designate them. Furthermore, no insight about their interaction can be inferred. GI values help. For singletons, the significant GI values, known to range in $[0, 1]$, are those of features $A1$ and $A2$. $GI(A2) > GI(A1)$ means that the evidence brought by $A2$ is greater than the one by $A1$, this is in agreement with the data. Only one pair is given a significant GI value, known to range in $[-1, 1]$, $A1$-$A2$. The negative sign characterizes this subset as redundant. For subsets of cardinality three, the GI shows some type of interaction in $A1$-$A2$-$A5$. This is in agreement with the residual correlation between $A5$-$A1$ and $A5$-$A2$ regarding class 1.

FM set	*A1*	*A2*	*A5*	*A6*	*A1-A2*	*A1-A5*	*A1-A6*	*A2-A5*	*A2-A6*	*A5-A6*	*A1-A2-A5*	*A1-A2-A6*	*A1-A5-A6*	*A2-A5-A6*
FM Value	0.45	0.93	0	0	1	0.72	0.45	0.93	0.93	0	1	1	0.72	0.93
GI	0.30	0.64	0.05	0	-0.52	0.13	0	-0.13	0	0	-0.27	0	0	0.0

Table 5: *Red* dataset: Fuzzy measure (FM) coefficients and GI values

Results on the *Comp* dataset are shown in Table 6. Fuzzy measure coefficients are one only for all subsets containing the complementary pair $A3$-$A4$ and zero for the remaining subsets including singletons $A3$ and $A4$ since neither $A3$ nor $A4$ can be individually used to identify class 1. GI values are the same for both singletons, indicating their equal relevance, and it is also one

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

for the pair. This positive interaction characterizes the complementariness of the two features.

FM set	*A3*	*A4*	*A5*	*A6*	*A3-A4*	*A3-A5*	*A3-A6*	*A4-A5*	*A4-A6*	*A5-A6*	*A3-A4-A5*	*A3-A4-A6*	*A3-A5-A6*	*A4-A5-A6*
FM Value	0	0	0	0	1	0	0	0	0	0	1	1	0	0
GI	0.5	0.5	0	0	1	0	0	0	0	0	0	0	0	0

Table 6: *Comp* dataset: Fuzzy measure (FM) coefficients and GI values

Above mentioned results confirm that the GI is positive valued for complementary feature pairs and negatively valued for complementary ones. In addition, they suggest that GI characterizes as relevant all subsets in it power set. These observations hold for higher cardinality. For instance, if subset $A1$-$A2$-$A3$ is complementary, GI also identify subsets $A1$-$A2$, $A1$-$A3$, $A2$-$A3$ as complementary, and the three singletons as relevant. Moreover, the sign of GI for both complementary and redundant sets of features of cardinality three was positive, suggesting an alternating behavior for redundant feature subsets.

These trends can be formalized in extreme situations, where a subset of features are fully redundant or complementary while the remaining only bring noise. The two following theorems are proved (the proof is given in the *Appendix*):

Theorem 1 - Complementary features

Let N be a dataset with n features among which, for a given class, c of them are fully complementary and $n-c$ are noise. Let C be the subset defined by the c fully complementary features. The GI for power set members of C with cardinality $a \in \{1, \ldots, c\}$ is:

$$GI(a) = \frac{1}{c - a + 1}$$

This yields $1/c$ for each of the c singletons and 1 for the whole feature subset C.

Theorem 2 - Redundant features

Let N be a dataset with n features among which, for a given class, r of them are fully redundant and $n - r$ are noise. Let R be the subset defined by the r fully redundant features. The GI for power set members of R with cardinality $a \in \{1, \ldots, r\}$ is:

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

$$GI(a) = \frac{(-1)^{a+1}}{r-a+1}$$

This yields $1/r$ for each of the r singletons, 1 for the whole feature subset R when r is odd and -1 when r is even.

The above results are used to propose some feature selection guidelines.

Feature selection guidelines using GI

Singletons: Relevant individual features are those with $GI > \frac{1}{n}$ [21].

Pairs: In this case, the union of complementary feature pairs, i.e., those with a significant positive GI, must be enriched with features coming from redundant feature pairs, i.e., those with a significant negative GI. This can be accomplished by checking that at least one of the features in each redundant feature pair is in the final set of selected features. A reasonable choice for GI thresholds is to set them to the mid value of the GI of fully complementary or redundant feature pairs already known to be ± 1, i.e., $GI \geq 0.5$ for the complementary case and $GI \leq -0.5$ for the redundant one. Alternatively, they can be set according to expert knowledge, as done in [14].

Higher cardinality sets: The pair selection reasoning can be extended to sets of higher cardinality. As follows from Theorems 1 and 2, the GI of fully complementary or redundant feature set of arbitrary cardinality is also ± 1 and thus, setting GI threshold to ± 0.5 remains valid.

To complete the classification process, once the Choquet integrals are computed for all the classes, using the selected features, the sample is assigned the class label, k, for which the support of the FM is maximum, as shown in Eq.(8):

$$k = \underset{k \in 1,\ldots,p}{arg\,max} \quad \mathcal{C}_{\mu^k} \tag{8}$$

6. Experimental results

The fuzzy measure and interaction index approach for feature selection and classification was evaluated on three UCI benchmark datasets (see Table

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

7). The Iris dataset is a well known dataset, easy to analyze, and allows for a fair comparison with results reported in [21].

	#Features	#Samples	#Classes
Iris	4	150	3
Breast Cancer	9	683	2
Wine	13	178	3

Table 7: Dataset characteristics

6.1. Experimental protocol

Datasets were first pre-processed with the raw data conversion method described in Section 4. Fuzzy measure coefficients required for the computation of the GI were estimated with the modified $HLMSr$ algorithm described in Section 3. For this purpose, the $HLMSr$ algorithm was set to work with a maximum of 2000 iterations and a learning rate $\alpha = 0.05$.

Classification performance was evaluated by means of the classification accuracy in 5-Fold cross validation experiments taking care that the original proportion of samples per class was preserved. Finally, the robustness CI classification under GI feature selection was also evaluated. For this purpose, features were ranked using the sum of their positions in the five GI ranks. This global ranking was then used to measure CI classification accuracy using feature subsets of increasing cardinality. GI feature selection results were compared with those obtained with the $SVM\text{-}RFE$ [29] feature selection technique based on a core SVM classifier set to work with a radial kernel and default constant complexity C=1. Features selected by the GI and $SVM\text{-}RFE$ techniques were used to evaluate the performance of SVM and CI classifiers. In these evaluations, the number of GI selected features was used as a cut-off for $SVM\text{-}RFE$.

6.2. Iris dataset

As shown in Table 8, fuzzy measure coefficients of $V1$, $V2$ and $V1$-$V2$ are null in all classes suggesting that they are irrelevant for the classification task. For class 1, the $GI = -1$ for coalition $V3$-$V4$ points out that these features are completely redundant, i.e., only one of them is needed. This is confirmed by the data projection in $V3$-$V4$, shown in Figure 3, and by the high relevance, $GI = 0.5$, of individual features $V3$ and $V4$.

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

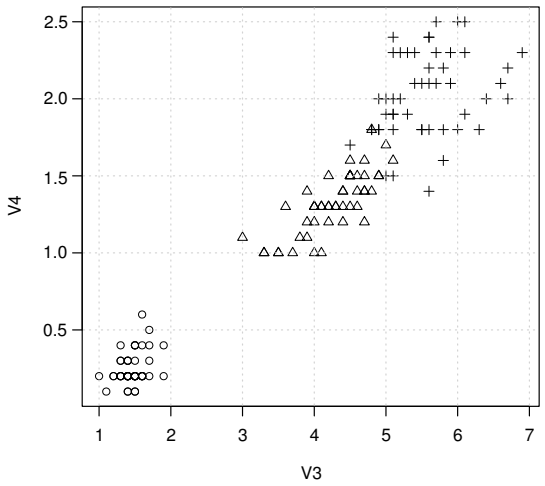

Figure 3: Iris data set. Class 1 is plotted with circles , Class 2 with triangles and Class 3 with plus sign

For class 3, the $GI = -0.56$ for coalition $V3$-$V4$ points out that these features are redundant for the class. For classes 2 and 3, the GI of feature $V4$ is higher than that of $V3$, pointing out that $V4$ is more relevant than $V3$ for these classes.

	Fuzzy measures			GI values		
Feature set	Class 1	Class 2	Class 3	Class 1	Class 2	Class 3
$V1$	0	0	0	0	0.01	0.01
$V2$	0	0	0	0	0	0
$V3$	1	0.21	0.61	0.5	0.29	0.33
$V4$	1	0.66	0.91	0.5	0.71	0.66
$V1$-$V2$	0	0	0	0	0	-0.01
$V1$-$V3$	1	0.26	0.63	0	0.02	0
$V1$-$V4$	1	0.66	0.96	0	-0.02	0.01
$V2$-$V3$	1	0.21	0.61	0	0	-0.01
$V2$-$V4$	1	0.66	0.96	0	0	0.02
$V3$-$V4$	1	1	1	-1	0.11	-0.56
$V1$-$V2$-$V3$	1	0.26	0.63	0	0	0.02
$V1$-$V2$-$V4$	1	0.66	0.96	0	0	-0.02
$V1$-$V3$-$V4$	1	1	1	0	0	-0.04
$V2$-$V3$-$V4$	1	1	1	0	0	-0.02

Table 8: *Iris* - Fuzzy measure coefficients of feature subsets together with their GI values

A comparison of these GI results with those reported in [21] shows that feature singletons are assigned roughly similar relevance values, except for $V1$

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

and $V2$ which relevance values are close to zero in our case. When considering feature pairs, the highest absolute GI value in [21] is met for coalition $V1$-$V3$ in class 2, a result quite difficult to explain. Similarly, coalition $V3 - V4$ in class 1 is assigned $GI = -0.05$ suggesting the lack of interaction between both features in this class, a result quite difficult to explain when observing the data.

Ranking	*SVM-RFE*					*GI*				
	P1	P2	P3	P4	P5	P1	P2	P3	P4	P5
1	V3	V4	V3	V4	V3	V4*	V4*	V4*	V4*	V4*
2	V4	V3	V4	V3	V4	V3*	V3*	V3*	V3*	V3*
3	V1	V1	V1	V1	V2	V1	V1	V1	V1	V1
4	V2	V2	V2	V2	V1	V2	V2	V2	V2	V2
# Misclassified by SVM	1	2	1	1	1	2	0	1	1	1
# Misclassified by CI	2	1	1	1	1	2	0	1	1	1

Table 9: *Iris* - Features rankings induced by the $SVM - RFE$ and the GI methods. Features relevant to at least one class in the GI method are marked with *. The number of misclassified samples by SVM and CI classifiers is shown at the bottom of each of partition, from P_1 to P_5.

GI and SVM-RFE feature selection results are shown in Table 9. In both methods, features $V3$ and $V4$ are more relevant than $V1$ and $V2$ for the classification. Although classification errors and selected features are roughly similar for both approaches, in agreement with the 96.7 % reported in [21] using all features, an interpretable feature characterization is additionally provided by the GI approach.

6.3. Breast Cancer dataset

This dataset has been widely used in scientific literature. To our knowledge, the best classification accuracy results have been obtained with IncNet (97.1%) [30]. With CI classifiers, the best reported classification accuracy is 91.5% [31]. As shown in Table 10, GI feature selection and CI classification yield a classification accuracy of 97.5%. Thanks to the untouched coefficient management in $HLMSr$ around 425 coefficients are identified on average over the different classes (512 for $HLMS$).

The four most significant features according SVM-RFE ($V1$, $V3$, $V6$ and $V7$) belong to set of relevant features selected by GI (Table 10). It is worth noting that GI selected features are also relevant for SVM as indicated by the reduction of misclassified samples. Stability of CI classification under

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

Ranking	*SVM-RFE*					*GI*				
	P1	P2	P3	P4	P5	P1	P2	P3	P4	P5
1	V3	V6	V3	V3	V6	V6*	V6*	V6*	V6*	V6*
2	V6	V7	V1	V1	V1	V3*	V1*	V7*	V3*	V3*
3	V1	V1	V6	V6	V4	V8	V3*	V1*	V1*	V7*
4	V7	V9	V8	V8	V7	V1	V7	V3*	V9*	V9*
5	V8	V4	V7	V9	V3	V2	V2	V2	V2	V4
6	V9	V8	V4	V7	V9	V5	V9	V8	V8	V8
7	V4	V2	V5	V4	V8	V9	V4	V4	V4	V1
8	V5	V3	V9	V5	V5	V7	V5	V9	V7	V5
9	V2	V5	V2	V2	V2	V4	V8	V5	V5	V2
# Misclassified by *SVM*	5	9	10	3	6	5	6	7	2	6
# Misclassified by *CI*	7	11	9	5	6	7	4	6	4	6

Table 10: *Breast Cancer* - Features rankings induced by the $SVM-RFE$ and the GI methods. Features relevant to at least one class in the GI method are marked with *. The number of misclassified samples by SVM and CI classifiers is shown at the bottom of each of partition, from P_1 to P_5.

GI feature selection with a global rank $\langle$ $V6$, $V1$, $V3$, $V7$, $V9$, $V2$, $V8$, $V4$, $V5$ $\rangle$ is shown Fig. 4. The classification error stabilizes after considering 7 out of 9 features.

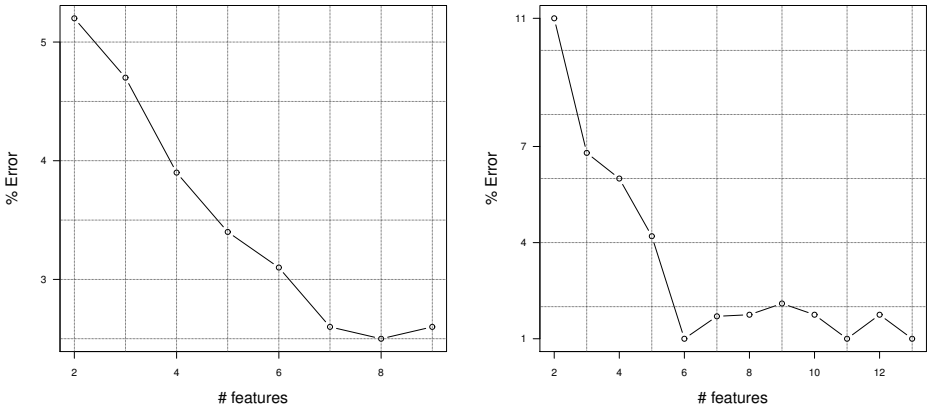

Figure 4: Classification error (%) of CI classifiers under GI feature selection with an increasing number of features in order of relevance for *Breast Cancer* dataset(left) and *Wine* dataset (right).

6.4. Wine dataset

As reported in [32], classes in the wine dataset are separable. However, only the regularized discriminant analysis method has been able to accom-

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

plish 100% classification accuracy. As shown in Table 11, using GI feature selection and CI classification, a classification accuracy of 98.5% can be accomplished. Thanks to modifications in $HLMSr$ relative to the treatment of untouched coefficients, on average over the different classes, around 840 coefficients are identified (8192 for $HLMS$).

Ranking	*SVM-RFE*					*GI*				
	P1	P2	P3	P4	P5	P1	P2	P3	P4	P5
1	V7	V7	V7	V7	V7	V13*	V13*	V13*	V13*	V13*
2	V11	V1	V1	V12	V1	V7*	V7*	V7 *	V7*	V10*
3	V1	V13	V4	V13	V12	V10*	V10*	V10*	V10*	V7*
4	V13	V12	V3	V1	V13	V11*	V11*	V11*	V12*	V11*
5	V12	V11	V12	V11	V4	V2*	V1*	V12*	V1*	V12*
6	V4	V3	V13	V4	V3	V1*	V12*	V1*	V11*	V1*
7	V3	V4	V11	V3	V10	V12*	V2*	V3*	V2	V2
8	V2	V10	V9	V10	V9	V4	V4	V2	V3	V4
9	V10	V8	V10	V2	V2	V6	V5	V6	V5	V5
10	V6	V6	V6	V6	V11	V5	V6	V4	V4	V6
11	V8	V2	V2	V9	V5	V3	V3	V5	V6	V3
12	V9	V5	V8	V8	V6	V9	V9	V9	V9	V9
13	V5	V9	V5	V5	V8	V8	V8	V8	V8	V8
# Misclassified by *SVM*	1	1	0	2	1	1	1	0	1	1
# Misclassified by *CI*	1	4	1	2	3	1	3	0	0	3

Table 11: *Wine* - Features rankings induced by the $SVM-RFE$ and the GI methods. Features relevant to at least one class in the GI method are marked with *. The number of misclassified samples by SVM and CI classifiers is shown at the bottom of each of partition, from P_1 to P_5.

The four most significant features according SVM-RFE, ($V7$, $V1$, $V12$ and $V13$) belong to set of relevant features selected by GI. As with *Breast cancer*, GI selected features are also good for SVM classification but the converse does not hold. Stability of CI classification under GI feature selection with a global rank $\langle$ $V13$, $V7$, $V10$, $V11$, $V12$, $V1$, $V2$, $V4$, $V3$, $V5$, $V6$, $V9$, $V8$ $\rangle$ is shown Fig. 4. The classification error stabilizes after considering 6 out of 13 features. This suggests that the CI classifier is robust to noise, since adding irrelevant features does not affect the classification error.

7. Conclusions

A computational strategy for the selection of relevant features based on feature interactions has been presented. The strategy is based on the careful characterization of feature interactions by means of fuzzy measures and the posterior computation of the generalized Shapley index, called GI. A detailed analysis of the raw data conversion process and the $HLMSr$ algorithm required for the identification of fuzzy measures revealed that both

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

tasks needed to be improved. Regarding the raw data conversion process, a normalization step that takes into account the distribution of features across all classes was introduced. The computational complexity of the $HLMSr$ algorithm was remarkably improved by disregarding monotonicity checking of untouched coefficients during algorithm's evolution. A more reliable estimation of fuzzy measure coefficients has been gained by identifying and solving the $HLMSr$ sensitivity to equal valued features.

A detailed analysis of the GI index on synthetic datasets containing noisy together with either redundant or complementary features uncovered a surprising GI behavior. The GI turns to be always positive in datasets with complementary features and of alternating sign in those with redundant ones. This result was used to propose some feature selection guidelines. Experimental results on benchmark datasets involving a dozen of features showed that competitive and interpretable fuzzy integral classifiers can be designed in this way. Furthermore, selected features seem to be also useful for the design of accurate SVM classifiers.

Further work is still needed to allow interpretable feature selection in datasets involving hundreds of features. In this regard, studies are first required to investigate the GI behavior involving both complementary and redundant features. Moreover, the reduction of both space and time required by the fuzzy measure identification to make the process tractable with huge datasets is still an open challenge.

8. Appendix

Mathematical formalization of features characterization for complementary and redundant features using GI

In extreme situations, when the only relevant features are either redundant or complementary, the GI expressions are easy to compute.

Theorem 1 - Complementary features. Let N be a dataset with n features among which, for a given class, c of them are fully complementary and $n-c$ are noise. Let C be the subset defined by the c fully complementary features. The GI for power set members of C with cardinality $a \in \{1, \ldots, c\}$ is:

$$GI(a) = \frac{1}{c - a + 1}$$

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

This yields $1/c$ for each of the c singletons and 1 for the whole feature subset C.

Proof: The set N of n features includes a set $C \subset N$ of c fully complementary features, while the remaining, $N \setminus C$, only bring noise. In this extreme situation, the only coefficients different from zero are the ones of C and the ones needed to keep monotonicity, i.e., the coefficients in $C \cup L$ with $L \in \mathcal{P}(N \setminus C)$. The value of all these coefficients is 1, as shown in Table 6 for $c = 2$. For given $A \subseteq N$, its GI is given by the Eq.(4):

$$I(A) = \sum_{K \subseteq N \setminus A} \frac{(n-k-a)!k!}{(n-a+1)!} \sum_{B \subseteq A} (-1)^{a-b} \mu(K \cup B) \tag{9}$$

Two cases have to be considered. The first case deals with subsets A entirely contained in C. The second case deals with subsets A having non-null fuzzy measure values only due to the monotonicity condition, i.e., $A = C \cup L$ with $L \in \mathcal{P}(N \setminus C)$.

- $A \subseteq C$, $a \leq c$

  In this case, B should be equal to A, otherwise, elements of C would be out of $K \cup B$, and the resulting coefficients are 0. As a result, $a - b = 0$ and $(-1)^{a-b}$ is always positive. The non null coefficients are those of the combinations for which $C \subseteq K \cup A$, meaning all the ones included in $K = \{C \setminus A\} \cup S$ where $S \subseteq N \setminus C$. As a result, $k = c - a + s$, $\mu(K \cup B) = 1$ and the Eq.(9) becomes:

$$I(A) = \sum_{S \subseteq N \setminus C} \frac{(n-c-s)!(c-a+s)!}{(n-a+1)!} \tag{10}$$

  This comes to:

$$I(A) = \frac{1}{(n-a+1)!} \sum_{s=0}^{n-c} \binom{n-c}{s} (n-c-s)!(c-a+s)! \tag{11}$$

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

As $\binom{n-c}{s} = \frac{(n-c)!}{s!(n-c-s)!}$:

$$I(A) = \frac{(n-c)!}{(n-a+1)!} \sum_{s=0}^{n-c} \frac{(c-a+s)!}{s!} \tag{12}$$

Remarking that $\frac{(c-a+s)!}{s!} = \binom{c-a+s}{s}(c-a)!$, it becomes:

$$I(A) = \frac{(n-c)!(c-a)!}{(n-a+1)!} \sum_{s=0}^{n-c} \binom{c-a+s}{s} \tag{13}$$

It can be proven, by recurrence, that[2]: $\sum_{s=0}^{n} \binom{m+s}{s} = \binom{m+n+1}{n}$, so replacing $\sum_{s=0}^{n-c} \binom{c-a+s}{s}$ by $\binom{n-a+1}{n-c}$ finally yields:

$$I(A) = \frac{1}{c-a+1} \tag{14}$$

- $C \subset A$, $a > c$

  All the sets B to consider must include C, they are of size $b = c + p$. There are $\binom{a-c}{p}$ of size $c + p$ in A which include C. The GI formula can be written as:

$$I(A) = \frac{1}{(n-a+1)!} \sum_{s=0}^{n-a} \left((n-a-s)!s! \sum_{p=0}^{a-c} \binom{a-c}{p} (-1)^{a-c-p} \right) \tag{15}$$

  The second sum is the binomial formula:
  $\sum_{p=0}^{n} \binom{n}{p} (-1)^{n-p} = \big(1 + (-1)\big)^n = 0.$

[2]Recurrence relation: $\sum_{s=0}^{n+1} \binom{m+s}{s} = \binom{m+n+1}{n} + \binom{m+n+1}{n+1} = \binom{m+n+2}{n+1}$

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

Hence, the only non-zero GI index values for a set N of n features where c of them are fully complementary and $n-c$ are noise are those belonging to subsets A in $\mathcal{P}(C)$. Furthermore, GI values depend just on the cardinality a of such subsets: $GI(A) = \dfrac{1}{c-a+1}$.

Theorem 2 - Redundant features. Let N be a dataset with n features among which, for a given class, r of them are fully redundant and $n-r$ are noise. Let R be the subset defined by the r fully redundant features. The GI for power set members of R with cardinality $a \in \{1, \ldots, r\}$ is:

$$GI(a) = \frac{(-1)^{a+1}}{r-a+1}$$

This yields $1/r$ for each of the r singletons, 1 for the whole feature subset R when r is odd and -1 when r is even.

Proof: The set N of n features includes a set $R \subset N$ of r fully redundant features, while the remaining , $N \setminus R$, only bring noise. In this extreme situation, all the coefficients belonging to $\mathcal{P}(N \setminus R)$ are assigned zero while all the subsets including at least one element from R are assigned a value of 1.Table 5 illustrates this trend with 2 partially redundant features. For given $A \subseteq N$, its GI formula in Eq.(4) can be written as:

$$I(A) = \sum_{K \subseteq N \setminus A} \frac{(n-k-a)!k!}{(n-a+1)!} \sum_{B \subseteq A} (-1)^{a-b} \mu(K \cup B) \tag{16}$$

As previously, $\mu(K \cup B)$ is either 0 or 1. The number of elements to sum can be computed as the total number of combinations according to the cardinalities minus the number of null coefficients. $\mu(K \cup B)$ is null if and only if both B and K are included in $\mathcal{P}(N \setminus R)$.

The total number of combinations is given by:

$$TotC = \sum_{b=0}^{a} \binom{a}{b} (-1)^{a-b} \sum_{k=0}^{n-a} \binom{n-a}{k} \frac{(n-k-a)!k!}{(n-a+1)!} \tag{17}$$

Let's define A as an union:
$A = \{Q \subset R\} \cup \{L \subset N \setminus R\}$, with $a = q + l$, and $q \leq r$ and $l \leq n-r$.

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

The sum of null coefficients over all K is computed from the all combinations of $N \setminus R$ in $N \setminus A$. It is given by:

$$NulC = \sum_{k=0}^{n-r-l} \binom{n-r-l}{k} \frac{(n-k-a)!k!}{(n-a+1)!} \tag{18}$$

This sum is non null only for $B \subseteq A \subseteq \mathcal{P}(N \setminus R)$.
Two cases have to be considered:

- $A \subseteq R$, $l = 0$

  Substituting l by zero and developing the binomial coefficient, Eq.(18) becomes:

  $$NulC = \frac{(n-r)!}{(n-a+1)!} \sum_{k=0}^{n-r} \frac{(n-k-a)!}{(n-r-k)!} \tag{19}$$

  This is exactly Eq.(12) with $k = n - r - s$.

  So, when $l = 0$, $NulC = \dfrac{1}{r-a+1}$.

  The number of subsets of size $p \leq a$ in A is $\binom{a}{p}$, but the only subset for which this difference is non null is $B = \emptyset$. All the others partial sums are the same, with alternate signs. The final GI absolute value is $NulC$, and the sign depends on the cardinality, a:

  $$I(A) = (-1)^{a+1} \frac{1}{r-a+1} \tag{20}$$

- $l \neq 0$

  Two cases have to be considered. The first, trivial, one is $q = 0$, $A \subseteq N \setminus R$. In this case, all the subsets B give the same sum with alternate signs and the index is zero.

  The remaining case includes both kinds of features: redundant and noisy ones. The partial sums are the same for all $B \subseteq Q$, $TotC$, and it is also the same, $TotC - NulC$, for all $B \subseteq L$. Thanks to the alternate signs and the sum of the binomial coefficients, the overall index is zero.

Author-produced version of the article published in Fuzzy Sets and Systems, 2015, N°270, p. 74-89
The original publication is available at http://www.sciencedirect.com
http://dx.doi.org/10.1016/j.fss.2014.09.015

For a set of r fully redundant features, while the others only bring noise, the GI values are null for all the combinations except for the power set of the r features. In this case, its value depends on the cardinality of the subset, a and is: $(-1)^{a+1}\frac{1}{r-a+1}$.

Acknowledgments

The authors are grateful to Tewfik Sari (Irstea, UMR ITAP) for the recurrence relation used in the theorems' proof.

References

[1] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[2] M. Grabisch, J.-L. Marichal, R. Mesiar, E. Pap, Aggregation functions: Means, Information Sciences 181 (1) (2011) 1 – 22. doi:10.1016/j.ins.2010.08.043.

[3] L. I. Kuncheva, L. C. Jain, Nearest neighbor classifier: Simultaneous editing and feature selection, Pattern Recognition Letters 20 (11-13) (1999) 1149–1156. doi:10.1016/S0167-8655(99)00082-3.

[4] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, J. P. Mesirov, Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles, Proceedings of the National Academy of Sciences of the United States of America 102 (43) (2005) 15545–15550.

[5] S. F. da Silva, M. X. Ribeiro, J. do E.S. Batista Neto, C. Traina-Jr., A. J. Traina, Improving the ranking quality of medical image retrieval using a genetic feature selection method, Decision Support Systems 51 (4) (2011) 810–820. doi:10.1016/j.dss.2011.01.015.
URL http://www.sciencedirect.com/science/article/pii/S0167923611000443

[6] H. Peng, F. Long, C. Ding, Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy, IEEE transactions on pattern analysis and machine intelligence 27 (8) (2005) 1226–1238.

[7] D. Denneberg, Non-Additive Measure and Integral, Kluwer Academic, 1994.

[8] M. Grabisch, H. Nguyen, E. Walker, Fundamentals of Uncertainty Calculi, with Applications to Fuzzy Inference, Kluwer Academic, 1995.

[9] Z. Wang, G. J. Klir, Fuzzy measure theory, Plenum, 1992.

[10] W. Wang, Z. Wang, G. J. Klir, Genetic algorithms for determining fuzzy measures from data, J. Intell. Fuzzy Syst. 6 (2) (1998) 171–183.
URL http://dl.acm.org/citation.cfm?id=1316027.1316028

[11] Z. Wang, K. Leung, J. Wang, A genetic algorithm for determining nonadditive set functions in information fusion, Fuzzy Sets and Systems 102 (1999) 436–469.

[12] Z. Wang, K. L. Wong, J. Wang, A new type of nonlinear and the computational algorithm, Fuzzy Sets and Systems 112 (2000) 223–231.

[13] Z. Wang, G. J. Klir, W. Wang, Monotone set functions defined by choquet integral, Fuzzy Sets and Systems 81 (2) (1996) 241 – 250. doi:10.1016/0165-0114(95)00181-6.
URL http://www.sciencedirect.com/science/article/pii/0165011495001816

[14] L. Mikenina, H.-J. Zimmermann, Improved feature selection and classification by the 2-additive fuzzy measure, Fuzzy Sets and Systems 107 (2) (1999) 197 – 218. doi:10.1016/S0165-0114(98)00429-1.
URL http://www.sciencedirect.com/science/article/pii/S0165011498004291

[15] M. Grabisch, A new algorithm for identifying fuzzy measures and its application to pattern recognition, in: Fourth IEEE international conference on fuzzy systems, Yokohama, Japan, 1995, pp. 145–150.

[16] J.-L. Marichal, M. Roubens, Determination of weights of interacting criteria from a reference set, European Journal of Operational Research 124 (3) (2000) 641 – 650. doi:10.1016/S0377-2217(99)00182-4.
URL http://www.sciencedirect.com/science/article/pii/S0377221799001824

[17] J. Wang, Z. Wang, Detecting constructions of nonlinear integral systems from input-output data: an application of neural networks, in: Fuzzy Information Processing Society, 1996. NAFIPS., 1996

Biennial Conference of the North American, 1996, pp. 559–563. doi:10.1109/NAFIPS.1996.534796.

[18] W. Jia, W. Zhenyuan, Using neural networks to determine sugeno measures by statistics, Neural Networks 10 (1) (1997) 183–195. doi:http://dx.doi.org/10.1016/S0893-6080(96)00080-9.
URL http://www.sciencedirect.com/science/article/pii/S0893608096000809

[19] M. Grabisch, E. Raufaste, An empirical study of statistical properties of the choquet and sugeno integrals, Trans. Fuz Sys. 16 (4) (2008) 839–850. doi:10.1109/TFUZZ.2008.917295.
URL http://dx.doi.org/10.1109/TFUZZ.2008.917295

[20] J. Murillo, S. Guillaume, E. Tapia, P. Bulacio, Revised hlms: A useful algorithm for fuzzy measure identification, Information Fusion 14 (4) (2013) 532 – 540. doi:http://dx.doi.org/10.1016/j.inffus.2013.01.002.
URL http://www.sciencedirect.com/science/article/pii/S1566253513000183

[21] M. Grabisch, The representation of importance and interaction of features by fuzzy measures, Pattern Recognition Letters 17 (6) (1996) 567 – 575. doi:10.1016/0167-8655(96)00020-7.
URL http://www.sciencedirect.com/science/article/pii/0167865596000207

[22] M. Grabisch, I. Kojadinovic, P. Meyer, A review of methods for capacity identification in choquet integral based multi-attribute utility theory: Applications of the kappalab r package, European Journal of Operational Research 186(2) (2008) 766–785.

[23] M. Grabisch, k-order additive discrete fuzzy measures and their representation, Fuzzy Sets and Systems 92 (2) (1997) 167–189. doi:10.1016/S0165-0114(97)00168-1.

[24] V. Torra, Y. Narukawa, Modeling Decisions, Springer, 2007.

[25] L. Shapley, A value for n-person games, in: H. Kuhn, A. Tucker (Eds.), Contributions to the Theory of Games, vol II, Vol. 28 of Annals of Mathematics Studies, Princeton University Press, 1953, pp. 307–317.

[26] T. Murofushi, S. Soneda, Techniques for reading fuzzy measures (iii): interaction index, in: 9th Fuzzy System Symposium, Sapporo, Japan, 1993, pp. 693–696.

[27] M. Grabisch, C. Labreuche, J.-C. Vansnick, On the extension of pseudo-boolean functions for the aggregation of interacting criteria, European Journal of Operational Research 148 (1) (2003) 28 – 47. doi:http://dx.doi.org/10.1016/S0377-2217(02)00354-5.
URL http://www.sciencedirect.com/science/article/pii/S0377221702003545

[28] S. Jullien, L. Valet, G. Mauris, P. Bolon, S. Teyssier, An attribute fusion system based on the choquet integral to evaluate the quality of composite parts, Instrumentation and Measurement, IEEE Transactions on 57 (4) (2008) 755–762. doi:10.1109/TIM.2007.913719.

[29] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, Machine Learning 46 (1-3) (2002) 389–422.
URL http://citeseer.ist.psu.edu/guyon02gene.html

[30] W. Duch, R. Adamczak, K. Grabczewski, A new methodology of extraction, optimization and application of crisp and fuzzy logical rules, Neural Networks, IEEE Transactions on 12 (2) (2001) 277–306. doi:10.1109/72.914524.

[31] S. Yue, P. Li, Z. Yin, Parameter estimation for choquet fuzzy integral based on takagi-sugeno fuzzy model, Information Fusion 6 (2) (2005) 175 – 182. doi:10.1016/j.inffus.2004.11.002.

[32] S. Aeberhard, D. Coomans, O. de Vel, Comparison of classifiers in high dimensional settings, Tech. Rep. 92-02, Dept. of Computer Science and Dept. of Mathematics and Statistics, James Cook University of North Queensland (1992).